Санкт-Петербургский государственный университет
Факультет прикладной математики - процессов управления
Кафедра технологии программирования

Шилов Илья Михайлович

Выпускная квалификационная работа бакалавра

Автоматическое выявление и расшифровка аббревиатур и сокращений в тексте

010400
Прикладная математика и информатика

Заведующий кафедрой,
кандидат физ.-мат. наук,
доцент
Сергеев С.Л.

Научный руководитель,
кандидат физ.-мат. наук,
доцент
Добрынин В.Ю.

Рецензент,
кандидат технических наук,
доцент
Гришкин В.М.

Санкт-Петербург
2016

Содержание

Принятые сокращения и обозначения

В данном списке перечисляются сокращения, используемые в работе:

LDA - Latent Dirichlet Allocation (латентное размещение Дирихле);

CBOW - Continuous Bag of Words;

Bag-of-words - мешок слов.

Dir - распределение Дирихле

$Milt$ - мультиномиальное распределение

РОМИП - Российский семинар по оценке методов информационного поиска;

$\propto$ - символ пропорциональности.

Введение

В современном мире, возникает все больше документации в электронном виде. Отчеты, договоры и другие виды документов стало гораздо проще и удобнее хранить. Все это представляет интерес для анализа. Объемы данных не позволяют проводить этот анализ вручную, поэтому разрабатываются алгоритмы и методы для машинного решения этой задачи.

Однако многогранность и сложность естественных языков, делает извлечение информации из текстов нетривиальной и трудоемкой работой, требующей предварительной обработки. Возникает целый спектр задач по предобработке документов. Например, удаление знаков препинания, изменение форм слов, приведение их к определенному виду. В частности, одной из таких проблем является раскрытие сокращений. В тексте возможны ситуации, когда слово встречается некоторое количество раз в своей полной форме, а далее применяется в сокращенной, подразумевая, что читатель сможет по контексту определить каждое слово. Для корректного анализа текстов требуется научиться раскрывать сокращения. Также решение этой задачи может быть полезно при чтении текста пользователем, в случае, если ему понадобиться помощь в понимании сокращений.

Глава 1

О задаче

1.1. Постановка задачи

Целью является анализ текстов, заключающийся в умении находить и верно раскрывать сокращения слов и аббревиатур. Для достижения этой цели были поставлены следующие задачи.

Во-первых, рассмотреть возможные способы нахождения сокращений в тексте. Этому может способствовать выявление орфографических и семантических особенностей слов. Орфографические особенности могут заключаться в том, что сокращения заканчиваются точкой, а аббревиатуры зачастую пишутся заглавными буквами. Семантические особенности - это смысловые значения, то есть контекст слов и темы, к которым они относятся.

Во-вторых, изучить методы, способствующие нахождению в тексте связей между словами. Эти связи, в свою очередь, могу помочь соотнести друг с другом полную и сокращенную форму слов.

1.2. Сокращения

Сокращения - это укороченная форма слов или словосочетаний. Они имеют несколько классификаций [11]. Отталкиваясь от этих классификаций возможно выявление орфографических особенностей.

Одна из классификаций основана на способах сокращения и выглядит следующим образом:

1. Графические сокращения - слова в которых часть букв в конце заменена точкой или часть букв в середине заменены дефисом. Например: рус. – русский, про-во – производство.
2. Инициальные аббревиатуры – слова, образованные из названий начальных букв или из начальных звуков слов, входящих в исходное слово-

сочетание.

3. Сложносокращенные слова – сокращенные слова, образованные из слов исходного сокращения, все или часть из которых были усечены. Например: колхоз – коллективное хозяйство, авиазавод – авиационный завод.

4. Высекаемые слова – слова, в которых высечены буквы, кроме начальных и конечных букв, а оставшиеся стянуты в сокращенное слово. Например: млрд – миллиард, млн – миллион.

5. Смешанные сокращения – сокращения, в которых сочетаются несколько способов образования сокращения. Например: НИИполиграфмаш – научно-исследовательский институт полиграфического машиностроения, кф. – кинофильм, стб. – столбец.

Другая классифицирует по степени распространенности:

1. Общепринятые – сокращения, употребляемые во всех изданиях.

2. Специальные – употребляются в изданиях отраслевой литературы для специалистов, принятые в этой отрасли понятные подготовленному читателю без расшифровки.

3. Индивидуальные сокращения – употребляются только в тексте одного издания, введенные его авторами или издателем и так или иначе расшифрованные в нем.

1.3. Корпус документов

В качестве обрабатываемого корпуса документов используется коллекция нормативных документов 2007 года семинара РОМИП. В нем содержатся тексты документов Законодательства РФ, Москвы и Санкт - Петербурга по состоянию на декабрь 2006.

1.4. Анализ задачи

Для нахождения возможных сокращений необходимо рассмотреть упомянутые выше орфографические и семантические особенности слов.

При поиске первых из них, проходя по тексту, будем запоминать слова, имеющие дефис или точку в конце, а также слова, у которых несколько букв написаны заглавными. Удобным инструментом для нахождения таких слов являются регулярные выражения.

Для анализа семантических особенностей, то есть смысловых, в документах будем подсчитывать различные статистики, такие как – количество вхождений слов в документ и частоты встреч сочетаний слов. На

основе этих данных, будет возможно выдвижение гипотез о соответствии их свойств реальным отношениям между словами.

Так будем предполагать, что слова, слишком редко встречающиеся в документах, не являются сокращениями, потому что сокращать слова имеет смысл при их частом употреблении для удобства записи и сокращения объема текста. Ко всему прочему о редких словах недостаточно статистической информации для выделения семантических особенностей, поэтому их можно не учитывать.

Заметим, что аббревиатуры в полной форме - это словосочетания. Вследствие чего, для нахождения соответствия между полной и сокращенной формой необходимо обладать достаточным количеством статистических данных. Имея такие данные, появляется возможность выделить сильно выраженные частоты соседства. Благодаря чему можем предположить, что если слово имеет небольшой набор соседствующих слов с повышенными частотами, то его сочетание с этими соседями возможно является полной формой аббревиатуры.

Расширяя набор гипотез об информации, находящейся в статистике о частотах соседства слов, будем считать, что слово достаточно полно определяется своими соседями. Тогда самые общеупотребимые сокращения (р. – река, г – город, ул. – улица) встречаются в очень большом количестве различных тем и контекстов. Под контекстом будем поимать набор слов в данном предложении. В то время как тема является одной из идей рассматриваемых в документе. При этом такие общепринятые сокращения используются в меньшем количестве тем и контекстов, чем слова, не имеющие никакой специализации, такие как предлоги, союзы, местоимения. Однако количество тем и контекстов слов может варьироваться в зависимости от их семантических значений. Таким образом, в частотном векторе соседей указанных выше общеупотребимых сокращений слов значения более равномерно распределены, чем у сильно специализированных слов. Вектор частот соседей сокращений более узкого характера может иметь особенности. Для оценки распределения частот и упомянутых особенностей будем использовать энтропию вектора частот соседей. Она отражает отсутсвие информации, содержащейся в векторе частот. Так, если слово является часто встречающимся соседом для большого количества слов, его энтропия высока, потому что одинаковые частоты большого количества соседей дают нам мало информации.

Для нахождения связей между словами рассмотрим несколько моделей представления текста. В одной из моделей каждому документу сопоставляется некоторый набор тем, используемых в нем. Предполагаем, что каждое слово может принадлежать теме с некоторой вероятностью. При этом обычно в текстах каждое слово используется в определенном количестве тем. Будем считать, что тема определяется набором слов, которые

в ней используются. Если удастся найти распределение тем по документам и слов по темам, то можно считать, что слова семантически связаны общей темой. Таким образом, будем предполагать, что слова в полной и сокращенной формах используются в одних и тех же темах. И для нахождения полной формы, соответствующей сокращению, можно опираться на слова близкие по набору тем, в которых они используются. Текст в такой модели рассматривается как "мешок слов", что подразумевает под собой извлечение информации о количестве слов для нахождения тем, не учитывая контексты. При анализе важны не столько положения слов, сколько их частоты. Можно сказать, что такая модель ищет связи на глобальном уровне. [2]

Другая модель рассматривает текст с локальной точки зрения, представляя его как набор субпредложений. Под этим понятием подразумевается английский термин sub-sentences, определяемый как текстовая единица, под которой может пониматься как обычное предложение, так и целые абзацы или главы. В каждом субпредложении соседями считаются слова, находящиеся в нем, независимо от расстояний между ними. Далее для каждого слова подсчитываются статистики количества соседств. По этим данным возможно предположение о семантических связях слов. Так, например, можно использовать нейросеть, которая обучится предсказывать вероятных соседей интересующего слова или само слово по заданным соседям. Тогда в качестве связи, по которой возможно нахождение слов сокращений, примем вероятность присутствия слов в одном субпредложении. В такой модели слова в полной и сокращенной форме должны часто соседствовать с одними и теми же словами.

Рассмотрим описанные модели на примере предложения:
"Утвердить откорректированные Москомархитектурой ГУП НИиПИ Генплана г Москвы"

В этом предложении все слова считаются соседними. Они же определяют контекст данного предложения. На примере сокращения НИиПИ (Научно-исследовательский и проектный институт градостроительства) рассмотрим возможные выводы сделанные на основе применения двух приведенных моделей. Слова в предложении определяют контекст использования сокращения и учитываются как его соседи. В данном случае это, например, "Генплан" или "ГУП". На основе этих соседей с помощью модели, рассматривающей текст с локальной точки зрения, можем предсказать слова, которые могли бы быть на месте "НИиПИ", исходя из контеста. Эти слова являются кандидатами на полную форму этого сокращения. Также слова из предложения учитываются в другой модели, которая рассматрвает текст глобально с точки зрения распредления тем. Таким образом, находятся темы к которым относится сокращение "НИиПИ", и кандидатами на полную форму становятся слова участвующие в этих же темах.

В рассмотренных гипотезах предполагается, что полная и сокращенная форма семантически близки. Описанные модели представлены в подходах LDA и word2vec.

1.5. Обзор литературы

Описание базовых идей информационного поиска изложены в книге Manning C. “Introduction to information retrieval” [5]. В ней рассмотрен подход, в котором текст воспринимается как “мешок слов”, то есть не учитывается информация о точном расположении слов. Также представлены различные варианты нахождения статистик текста, включающих в себя подсчет количества вхождений слов в документы и частоту соседства слов.

Разработанный Mikolov T. подход word2vec детально описан в его статьях “Efficient Estimation of Word Representations in Vector Space” [6] и "Distributed Representations of Words and Phrases and their Compositionality" [7]. В этом подходе, основываясь на частотах использования, слова кодируются алгоритмом Хаффмана. Реализованы две архитектуры нейросети CBOW и Skip-gram, с помощью которых осуществляется предсказание слов по соседям и наоборот.

В исследовании “Latent Dirichlet Alloacation” David Blei [1] изучает модель LDA. В ней текст представляется в виде вероятностной смеси тем. А они, в свою очередь, состоят из вероятностной смеси слов. Также модель LDA рассмотрена в статье "Parameter estimation for text analysis" Heinrich G. [3]. В ней приведены детали реализации генеративной модели, свойства гиперпараметров, а также способы нахождения всех параметров по известному корпусу документов.

В статье “Deciphering Journal Abbreviations with JAbbr” Keith Jenkins [4] рассмотрена реализация нахождения полной формы для сокращенных названий журналов. Неизвестная аббревиатура, указанная пользователем, приводилась в формат регулярного выражения, которое предполагало возможный набор слов, начинающийся с указанных букв. После чего этим регулярным выражением проверялась вся библиотека журналов и находились возможные варианты.

В статье Shannon C. “A Mathematical Theory of Communication” [9] представлены исследования понятия энтропии как меры информации. Слова представленные векторами частот их соседей можно оценить с точки зрения этой величины. Эта характеристика показывает некую меру информации или ее отсутсвия в векторах. Интерпретация энтропии даст дополнительную оценку, является ли слово сокращением.

В лекциях Воронцова К. В. [10] рассмотрены представление текстов, а также особенности моделей LDA и word2vec.

В презентации Moody C. [8] рассмотрены особенности моделей word2vec и LDA. Также предложен гибридный алгоритм lda2vec.

Идея оценки качества полученных результатов была подчерпнута из статьи “Reading Tea Leaves: How Humans Interpret Topic Models” [2]. Так как корпус документов не размечен, то есть нет информации о том, какие слова точно являются сокращениями и какие у них полные формы, то автоматически оценивать верность подобранных соответсвий между словами сокращениями и их полными формами, а также автоматически оценивать верность выборки возможных сокращений из текста невозможно. Поэтому был применен подход экспертной оценки.

Глава 2

Подготовка данных для решения задачи

2.1. Чтение корпуса документов

Корпус документов имел усложняющие обработку знаки пунктуации, символ амперсанта, а также шумы в виде случайных символов английского языка. Для чтения и извлечения информации использовались регулярные выражения. Это механизм нахождения в тексте образцов подходящих под описание, которое осуществляется на специальном языке регулярных выражений.

При последовательном чтении документов был создан словарь уникальных слов. У каждого слова подсчитывалась частота его встречаемости в документах. Для того чтобы обеспечить уникальность элементов в словаре, при разбиении текста было проведено приведение слов к начальной форме и нижнему регистру.

Напрмер, “Москве!”, “Москва.” и “Москвы,” сохранялось как слово “москв”. Словарь отсортирован по частоте встречаемости, для учета редко встречающихся слов. В результате получен словарь, в котором каждому слову соответсвует его частота:

$$\begin{pmatrix} \text{москв} - 357115 \\ \text{правительств} - 136087 \\ \text{город} - 125116 \\ \text{фонд} - 31907 \\ \dots \end{pmatrix}$$

2.2. Создание матрицы соседства

Для анализа частот соседей создается матрица соседства. Извлекая слова из документов, для каждого из слов подсчитывается его вектор частот встречи соседних слов. Таким образом, при встречи рядом со словом

“г” слова “москв” увеличивалась компонента, отвечающая за частоту встречи слова “москв” вектора соседей слова “г”.
Так для слова “ниипи”, рассмотренного в пункте 1.4, вектор соседства, отсортированный по частотам выглядит следующим образом:

$$\begin{pmatrix} \text{предусмотрен} - 611 \\ \text{администрац} - 377 \\ \text{федерац} - 117 \\ \dots \end{pmatrix}$$

2.3. Bag-of-words

Для модели LDA корпус документов был представлен в виде “мешка слов”. В таком представлении каждый документ является вектором пар, состоящим из слова и его частоты использования в документе. Все слово индексируется. В следствие чего получается более удобная система хранения информации о документах в виде пар, образованных индексом и частотой встречаемости, соответствующих одному слову.

2.4. Представление текста для word2vec

Аналогично предыдущему подходу, в word2vec каждое слово индексируется и хранится вместе с частотой. Это необходимо для дальнейшего кодирования слов. При этом документ во время обработки рассматривается как набор предложений.

Глава 3

Поиск сокращений

3.1. Понятие энтропии

Одной из семантических особенностей слова будем считать количество контекстов, в которых оно упоминается. Контекст определяется несколькими соседствующими словами. Частоты соседей могут давать нам некое понимание о семантических особенностях слова. Опираясь на них будет возможно выделение интересующих слов из корпуса документов. В данном контексте под интересующими словами подразумеваются вероятные сокращения или слова точно ими не являющиеся. Благодаря этой информации мы можем избавиться от неспециализированных слов, уменьшая таким образом сложность выбора. Предположим, что есть слово, которое встречается в документах большое количество раз. При этом для него нельзя выделить никакую группу слов с ярко выраженной частотностью соседства. Тогда это слово с большей частью слов встречается одинаково среднее количество раз. Из чего можно сделать вывод, что такое слово используется в слишком большом спектре контекстов. Это означает, как было указано выше, отсутсвие его специализации. В таком случае семантической информации в частотном векторе очень мало.

Появляется необходимость выделить численную меру предоставляемой вектором частот соседей информации. Для этого введем понятие энтропии для оценки слов.

Информационная энтропия – мера неопределенности или непредсказуемости информации. Пусть есть независимое случайное событие x с n возможными состояниями. $p(i)$ - вероятность i-го состояния. Клод Шеннон предположил, что прирост информации равен утраченной неопределенности. Он задал требования к ее изменению. Было доказано, что единственная функция, удовлятворяющая этим требованиям, имеет вид:

$$H(x) = -\sum_{i=1}^{n} p(i) \cdot \log_2(p(i))$$

В нашем случае событие x - выбор соседа при фиксированном слове.

n исходов - это n слов, которые когда-либо были рядом с рассматриваемым словом. Вероятность $p(i)$ - вероятность соседства.

3.2. Матрица соседства

Для подсчета энтропии необходимо иметь для каждого слова вектор частот его соседей. Составим матрицу N, где частота встречи j-го слова рядом с i-м равна $N[i][j]$.

Введем вектор $NSum$. В $NSum[i]$ будет храниться сумма частот соседей i-го слова в матрице N

Тогда вероятность соседства $p(i)$ – это частота соседства i-го слова со словом x, разделенная на общее количество подсчитанных соседств:

$$p(i) = \frac{N[x][i]}{Nsum[x]}$$

Таким образом, энтропия слова x равна:

$$H(x) = -\sum_{i=1}^{n} \frac{N[x][i]}{Nsum[x]} \cdot \log_2(\frac{N[x][i]}{Nsum[x]})$$

Глава 4

Word2vec

Представим один из подходов, позволяющих после обучения проецировать слова в векторное пространство с семантической близостью.

Word2vec – это модель для расчета векторных представлений слов, который реализует две основные архитектуры: Continuous Bag of Words (CBOW) и Skip-gram. В качестве входных данных используется предобработанный корпус документов, после использования алгоритма получаем набор векторов слов.

4.1. Семантическая близость

Семантическая близость слов будет определяться как косинусная близость между векторами слов. Косинусная близость – это мера сходства между двумя векторами, которая измеряет косинус угла между ними:

$$similarity(a, b) = \cos \alpha(a, b) = \frac{a \cdot b}{\| a \| \cdot \| b \|}$$

Word2vec стремится максимизировать семантическую близость слов, которые появляются рядом друг с другом, и напротив минимизировать семантическую близость слов, не встречающихся рядом. “Рядом” в данном случае означает в одном предложении.

Например, слова “анализ” и “исследование” часто встречаются в похожих контекстах: “Ученые провели анализ алгоритмов” или “Ученые провели исследование алгоритмов”. В данном случае word2vec будет корректировать вектора этих дву слов так, чтобы максимизировать их косинусную близость, потому что они встречаются в схожих контекстах.

4.2. Принцип работы

В ходе обучения модели word2vec осуществляются следующие шаги:

1. Корпус документов прдеставляется в виде субпредложений, которые обрабатываются друг за другом. Рассчитывается встречаемость каждого слова в корпусе (то есть количество раз, когда слово встретилось в корпусе – и так для каждого слова).

2. Массив слов сортируется по частотам. Из него удаляются редкие слова. На основе пар, состоящих из слов и их частот, полученная выборка сохраняется в хэш-таблице.

3. Строится дерево Хаффмана для кодирования словаря, учитывая частоты встречаемых слов. Получаем для каждого слова код.

4. Из корпуса читается так называемые субпредложения и проводится субсэмплирование наиболее часто встречающихся слов. Субпредложение – это некий базовый элемент корпуса, чаще всего – стандартное предложение, но это может быть и абзац, и целая статья. Субсэмплирование – это процесс изъятия наиболее частотных слов из анализа, что ускоряет процесс обучения алгоритма и способствует значительному улучшению качества получающейся модели.

5. По субпредложению проходим окном (размер окна задается алгоритму в качестве параметра). В данном случае под окном подразумевается максимальная дистанция между текущим и предсказываемым словом в предложении. То есть, если окно равно трем, то для предложения “Утвердить откорректированные Москомархитектурой ГУП НИиПИ Генплана г Москвы” анализ (применение алгоритмов на базе одной из архитектур CBOW или Skip-gram) будет проходить внутри блоков в три слова: “Утвердить откорректированныеМоскомархитектурой”, “откорректированные Москомархитектурой ГУП” и так далее.

6. Применяется нейросеть прямого распространения с функцией активации иерархический софтмакс. Также иногда используется негативное сэмплирование для минимизации близости с семантически далекими словами. Иерархический софтмакс лучше подходит для работы с не очень частотными словами, работает медленнее негативного сэемплирования. А оно в свою очередь лучше подходит для работы со словами с большой частотой и векторами слов небольшой размерности (50-100), при этом работает быстрее.

4.3. Архитектуры

CBOW – архитектура нейронной сети, в которой корпус документов просматривается окном ширины $2 * h + 1$. Для каждого окна однослойная нейронная сеть предсказывает центральное слово $w(t)$ по окружающим его $w(t+1), i \in [-h; h] \setminus \{0\}$ словам. Данная архитектура представлена на рисунке 4.1, на котором отражен принцип работы, выраженный в генерации центрального слова используя h соседей слева и h соседей справа от него.

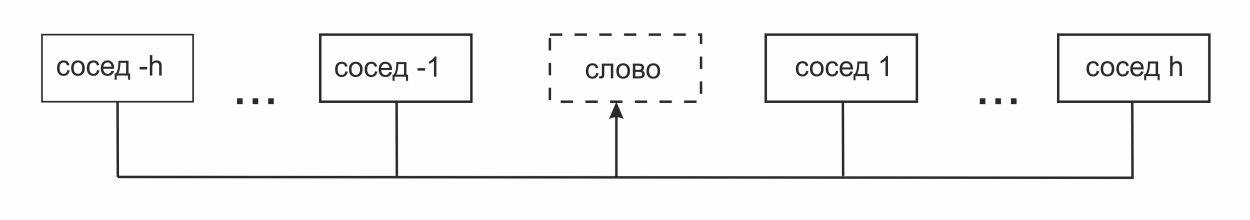

Рис. 4.1: Архитектуры CBOW

Skip-gram - архитектура нейронной сети, в которой корпус документов просматривается окном ширины $2 * h + 1$. Для каждого окна однослойная нейронная сеть предсказывает $2 * h$ окружающих слов $w(t+1), i \in [-h; h] \setminus \{0\}$ для данного центрального слова $w(t)$. Данная архитектура представлена на рисунке 4.2, на котором отражен принцип работы, выраженный в генерации h соседей слева и h соседей справа, используя центральное слово.

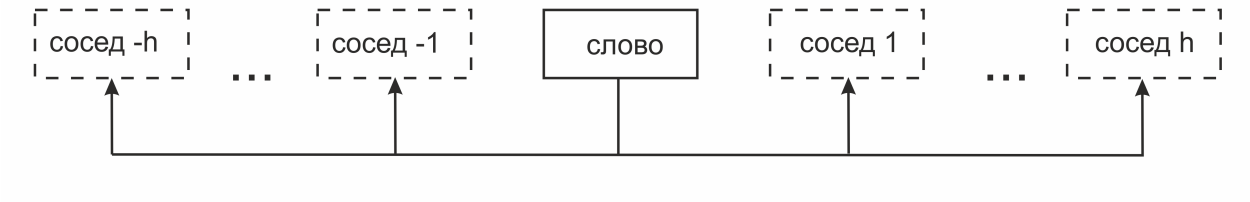

Рис. 4.2: Архитектура Skip-gram

Глава 5

LDA

LDA (Latent Dirichlet allocation) – генеративная модель. Она предполагает, что в корпусе каждый документ представляет из себя смесь неких тем. Каждой теме соответствует набор слов, которые принадлежат теме с разной вероятностью. Модель LDA может генерировать корпус документов, а также помогает объяснить, схожесть частей текста.

5.1. Алгоритм генерации

В модели задается количество тем K, количество документов M, гиперпараметры α и β для распределений Дирихле.

Для каждой темы k строим распределение φ_k – слов по теме k:

$$\varphi_k = Dir(\beta)$$

Для каждого документа m строим распределение ν_m – тем в этом документе:

$$\nu_m = Dir(\alpha)$$

Выбираем длины документов N_m. По очереди для каждой позиции от 1 до N_m выбираем индекс темы $z_{m,n}$ по мультиномиальному распределению относительно смеси тем ν_m в документе:

$$z_{m,n} = Mult(\nu_m)$$

По выбранному индексу темы выбирается слово $w_{m,n}$ из мультиномиального распределения слов в выбранной теме $\varphi_{z_{m,n}}$:

$$w_{m,n} = Mult(\varphi_{z_{m,n}})$$

5.2. Сэмплирование по Гиббсу

Мы хотим по корпусу документов построить модель LDA. То есть определить из текстов скрытые темы, их распределение по документам, а также распределение слов по темам. Сэмплирование по Гиббсу - это алгоритм генерации совместного распределения множества скрытых тем. Общий принцип заключается в выборе произвольной темы и изменению ее в зависимости от всех остальных. Выбранные темы у нас определяются индексами $z_{m,n}$.

Пусть:
$n_m^{(k)}$ - количество слов в документе m темы k;
n_m - количество слов в документе m;
$n_k^{(t)}$ - количество слов t в теме k;
n_k - количество слов в теме k;
В работе [3] описан следующий алгоритм сэмплирования по Гиббсу:

1. Обнуляем $n_m^{(k)}, n_m, n_k^{(t)}, n_k$
2. Для каждого документа m генерируем индексы тем $z_{m,n}$ для всех словопозиций. В соответствии с этим увеличиваем статистики $n_m^{(k)}, n_m, n_k^{(t)}, n_k$
3. Итеративно повторяем последовательность действий:

   В каждом документе m: для каждого слова n уменьшаем статистики $n_m^{(k)}, n_m, n_k^{(t)}, n_k$ на единицу, высчитываем новую тему $\overline{k}$, используя вероятности принадлежности слов к темам, для слова на позиции n в соответсвии с этим увеличиваем на единицу необходимые статистики.

Вероятность того, что i-е слово документа m, принадлежит теме k высчитывается по формуле:

$$p(z_i = k \mid \overrightarrow{z}_{\neg i}, \overrightarrow{w}) = \frac{p(\overrightarrow{w}, \overrightarrow{z})}{p(\overrightarrow{w}, \overrightarrow{z}_{\neg i})} = \frac{p(\overrightarrow{w} \mid \overrightarrow{z})}{p(\overrightarrow{w}_{\neg i} \mid \overrightarrow{z}_{\neg i}) \cdot p(w_i)} \cdot \frac{p(\overrightarrow{z})}{p(\overrightarrow{z}_{\neg i})}$$

$$\propto \frac{n_{k,\neg i}^{(t)} + \beta}{\sum_{t=1}^{V} (n_{k,\neg i}^{(t)} + \beta)} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K} (n_m^{(k)} + \alpha) - 1}$$

$$\propto \frac{n_{k,\neg i}^{(t)} + \beta}{\sum_{t=1}^{V} (n_{k,\neg i}^{(t)} + \beta)} \cdot (n_{m,\neg i}^{(k)} + \alpha)$$

$$\overrightarrow{w} = \{w_i = t, \overrightarrow{w}_{\neg i}\}, \overrightarrow{z} = \{z_i = k, \overrightarrow{z}_{\neg i}\}$$

где $\neg i$ - все индексы, кроме i

После выполнения данной процедуры происходит извлечение распределений:

1) Слов по темам:

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta}{\sum_{t=1}^{V}(n_k^{(t)} + \beta)}$$

2) Тем по документам:

$$\nu_{m,k} = \frac{n_m^{(k)} + \alpha}{\sum_{k=1}^{K}(n_m^{(k)} + \alpha) - 1}$$

Для визуализации алгоритма удобно использовать байесовскую сеть. Это графическая модель, представляющая из себя направленный граф. В его узлах находятся переменные: гиперпараметры, индексы тем и так далее (см. рисунок 5.1). Направленные ребра показывают зависимость между переменными. Так вершина, в которую приходит ребро зависит от вершины, из которой оно исходит. На рисунке 5.1 представлена байесовская есть для модели LDA.

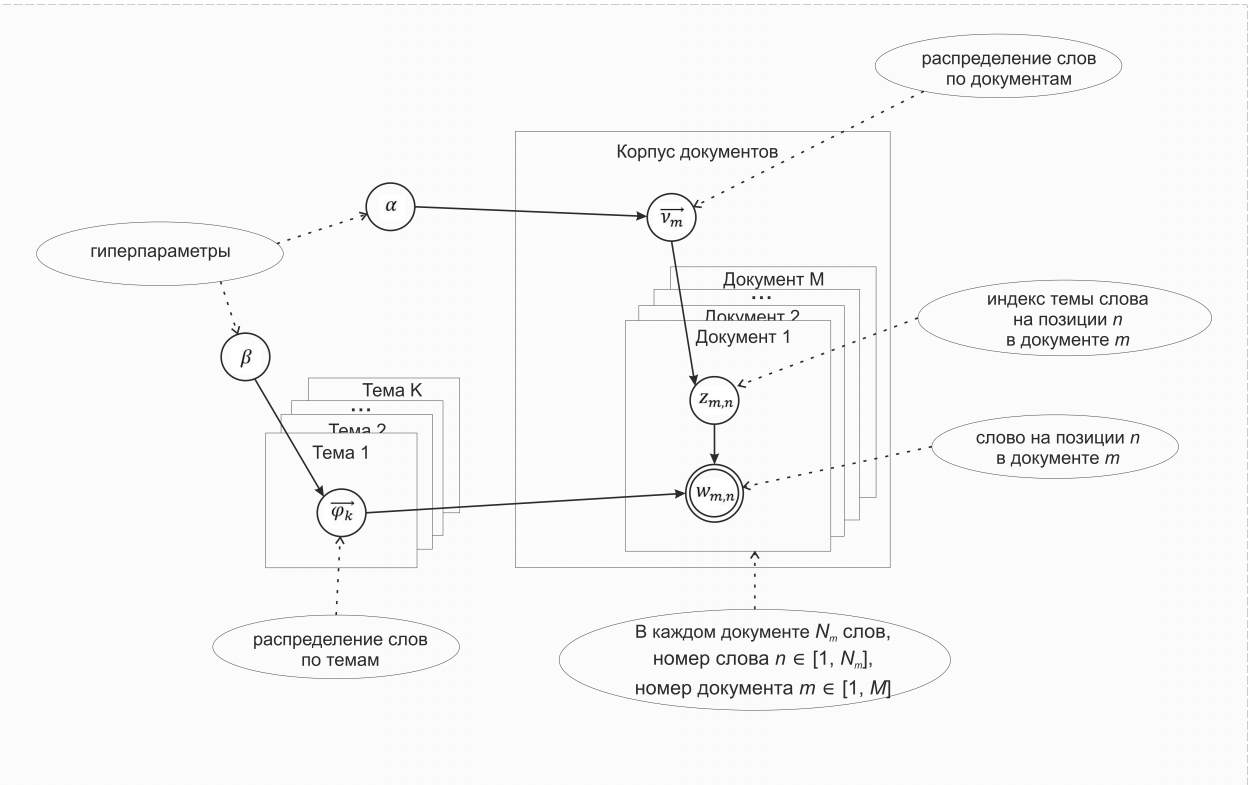

Рис. 5.1: Байесовская сеть модели LDA

Глава 6

LDA и word2vec

6.1 Отличия

У LDA и word2vec есть несколько отличий.

Первое отличие - это требуемый формат входные данных. В Word2vec при обучении текст передается субпредложениями. В то время как LDA принимает на вход документы, представленные в виде модели “мешок слов”.

Второе отличие заключается в подходе к предсказанию. В word2vec предсказывается слово по контексту близлежащих слов или наоборот. В LDA предсказывается тема слова исходя из статистик числа слов в документах и предположении о неких вероятностных распределениях тем в документах и слов по темам. Таким образом, word2vec анализирует локальную область текста, а LDA глобальную.

Третьим немаловажным отличием является полученное представление. Word2vec дает векторное пространство слов, в котором существует семантическая регулярность, то есть если из вектора мужчина вычесть вектор женщина и прибавить вектор слова королева, то получится вектор слова король. LDA дает распределение слов по темам, а это по сути является мягкой кластеризацией.

Четвертым отличием можно считать интрепретацию векторов слов. Типичный вектор слова в модели word2vec - это плотный вектор вещественных чисел
$[-0.75, -1.25, -0.55, -0.12, +2.2]$
В то время как, типичный вектора LDA - разряженный вектор вероятностей слов принадлежности к теме
$[0, 0.09, 0.78, 0.11]$
Разряженность вектора означает, что большая часть информации равна нулю, поэтому получаем большую интерпретируемость модели. В свою очередь, плотный вектор word2vec не может дать такую же степень интерпретируемости, зато очень выигрывает в гибкости, так как большая плотность вектора дает больше степеней свободы.

6.2 Совместное использование

Moody C. реализовал и представил гибридный алгоритм lda2vec. Для того чтобы иметь возможность предсказывать слово как локально, так и глобально автор к каждому вектору слова модели word2vec прибавляет разреженный вектор LDA. Затем к получившемуся в результате вектору можно прибавлять различные заранее известные категориальные признаки, после чего рассчитывать распределение слов по темам.

Таким образом, когда мы говорим о предсказании слов в тексте, то каждое слово предсказывается не только исходя из наиболее вероятного расположения в контексте, как в word2vec, но и исходя из вероятности слов встречи друг с другом в определенной теме и с определенными признаками.

lda2vec - это метод тематического моделирования аналогично LDA, но дополнительно использующий word2vec. В поставленной задаче добавление категориального признака осложнено тем, что в корпусе документов для каждого слова это по сути означает прикрепление ярлыка, что в ручном режиме не осуществимо, а в автоматическом необходимо разрабатывать дополнительный алгоритм определения этого категориального признака.

Глава 7

Программная реализация

7.1 Используемые инструменты

Для анализа корпуса документов была написана программа-прототип в интерактивной оболочке ***Jupyter*** языка программирования ***Python***. Были применены следующие пакеты:
re - анализ текста с помощью регулярных выражений;
gensim - библиотека для анализа больших корпусов документов и построения на их основе информационных моделей;
nltk - библиотека для анализа языковых особенностей слов, таких как приведение их к начальной форме;
collections - библиотека для составления словарей слов и их хранения.
lda2vec - реализация гибридного метода Moody C.

7.2 Описание обработанных данных

В ходе работы программы был проанализирован корпус из 100 документов длиной от 2950852 до 33288921 символов. Составлен словарь уникальных слов с их частотами длинной 55354, из которых чаще 100 раз используется 6464 слов. Построена модель “bag-of-words” для алгоритма LDA.

В итоге по этим данным были построены и обучены модели word2vec, LDA и lda2vec.

7.3 Описание процесса работы

В программе реализованы шаги представленные на схеме 7.1.

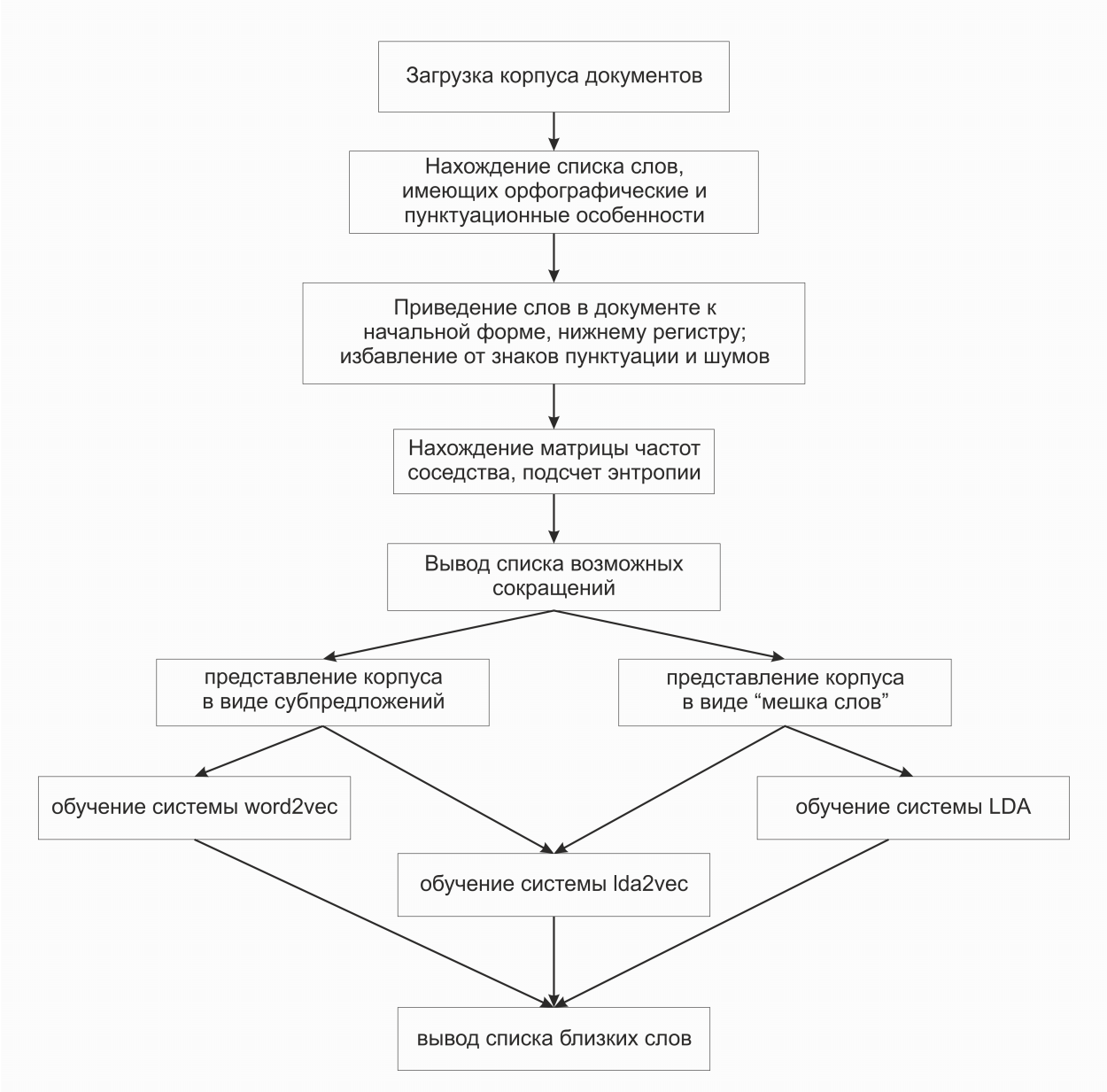

Рис. 7.1: Схема работы программы

Глава 8

Эксперименты

8.1. Поиск сокращений

По матрице частот соседства в пределах 5, 10, 15 и 20 элементов была посчитана энтропия. Гипотеза о большой энтропии у слов имеющих большое количество тем, и используемых в разнообразных контекстах подтвердилась. Слова "для, ряд, к, под" имеют энтропию больше 9.8, они показаны на графике 8.1.

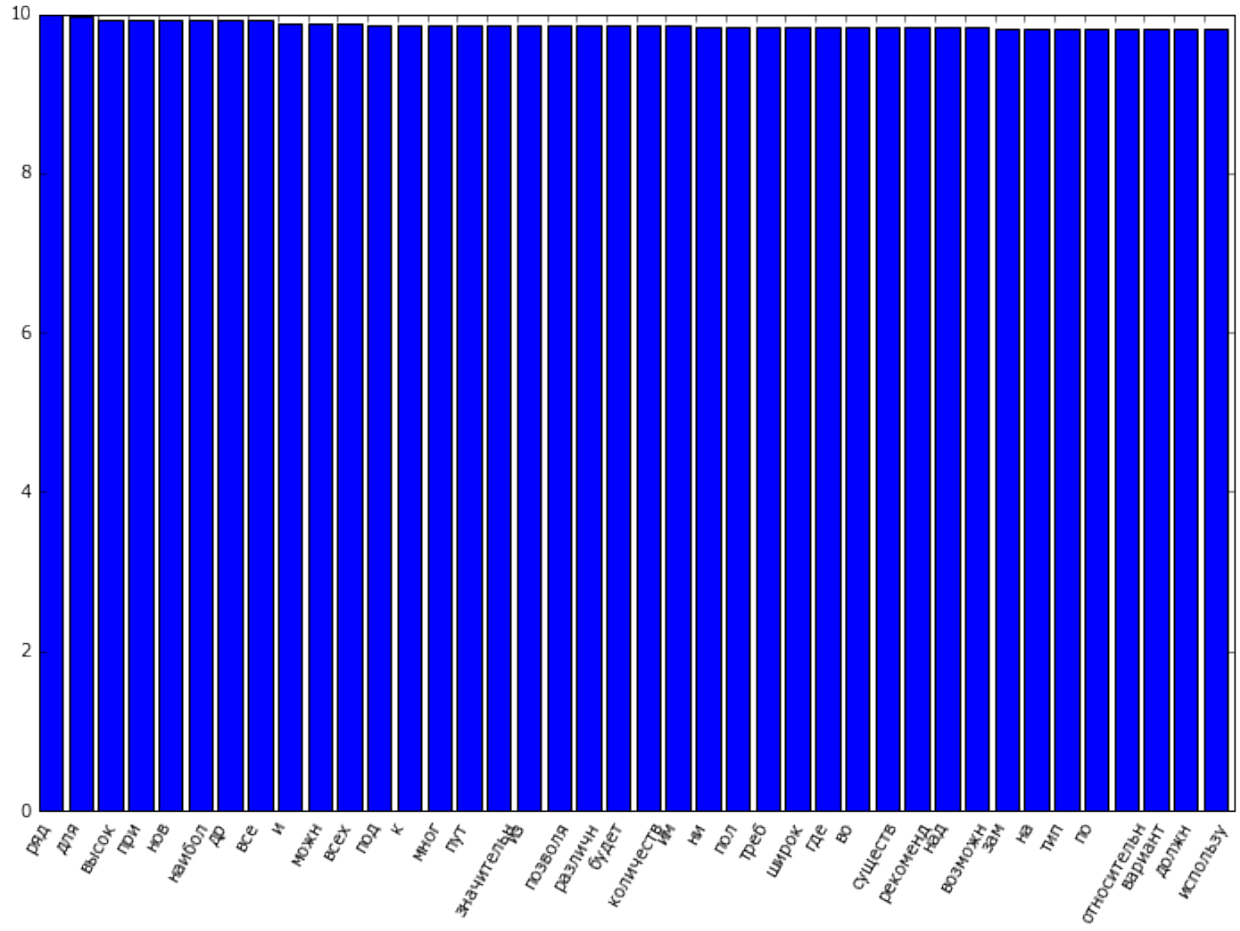

Рис. 8.1: Энтропия больше 9.8

Однако ни малое, ни среднее значение энтропии не дает оснований для однозначного определения принадлежности слова к сокращениями, потому что среди этих слов присутсвуют в одинковой мере как слова сокращения, так и слова ими не являющиеся. Слова с малым значением энтропии показаны на графике 8.2. Общее распределение энтропии можно увидеть в приложении на графике 8.3. Эта характеристика может помочь при отбрасывании слов с большой энтропией из списка возможных сокращений.

Список возможных сокращений получен с помощью регулярного выражения, предусматривающего слова содержащие точку в конце, дефис или имеющие больше одной заглавной буквы в написании.

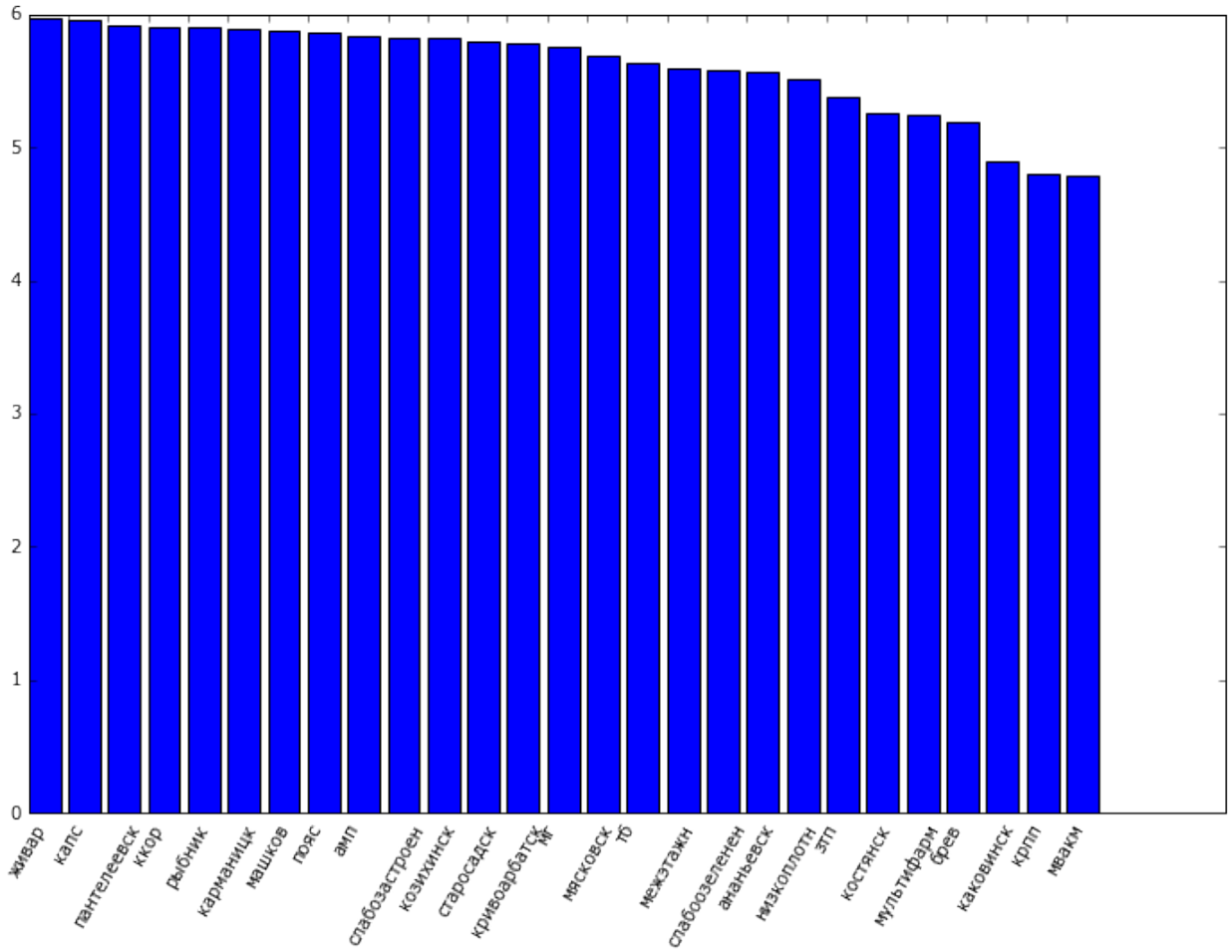

Рис. 8.2: Энтропия меньше 6

8.2. Поиск полной формы сокращения

По предложенному слову модель word2vec предлагает 20 наиболее семантически близких слов. В свою очередь модель LDA выбирает наиболее подходящую слову тему, а затем предлагает самые вероятные слова из нее. lda2vec производит аналогичные LDA действия.

8.3. Оценка качеств моделей

Корпус документов не имеет разметки, поэтому осуществление адекватной автоматизированной проверки верности расшифровки не представляется возможным. Таким образом, для оценивания качества подбора возможных полных форм сокращений была использована экспертная оценка. Данный подход был описан в статье “Reading Tea Leaves: How Humans Interpret Topic Models” [2]. В оценке приняло участие два эксперта.

Процесс оценивания был реализован следующим образом. В начале эксперты определяли, сколько из предложенных сокращений, действительно являлись сокращениями. Далее экспертам предлагалось оценить, есть ли среди 10 слов, предложенных каждой моделью, полная форма сокращения или ее часть в случае, если полная форма состоит из нескольких слов. А также может ли помочь данный список слов понять некую информацию о данном сокращении.

Среди 300 самых частых слов из списка возможных сокращений, не сокращениями оказалось 127. При этом больше 100, не подходящих слов, имеют длинну более 6 символов. В то время как, все сокращения имеет длинну не более 6 символов. На оставшихся данных наблюдалась аналогичная тенденция, связанная с количеством символов в слове. Также при снижении частоты используемых терминов, снижалось процентное содержание сокращений. Исходя из этого, будем в выборке возможных сокращений отбрасывать слова длинной более 6 символов.

По оценкам экспертов при использовании модели word2vec слова подходящие на роль полной формы или ее части редко присутсвовали в списке предложенных слов. При этом графические сокращения находились чаще. Это может быть связано с тем, что аббревиатуры по причине громоздкой записи полной формы, редко в ней используются. Поэтому по мнению экспертов, построенная модель показывает не столько полную форму, которая редко встречается, сколько контекст использования. Оценка экспертов результатов работы модели LDA и lda2vec схожа с предыдущей моделью. Так зачастую перечисленные слова, хорошо объединялись в темы, но слов дающих полную форму в них не присутствовало. Таким образом, они скорее помогали понять область использования данного сокращения, нежели его полную форму.

Выводы

Для обнаружения сокращений были выделены орфографические особенности на основе регулярных выражений, а также семантические на основе статистик соседства слов. По данным особенностям, с учетом выводов из экспертной оценки, можно сказать, что выделение сокращений из текста может быть достаточно успешно осуществлено.

Рассмотренные подходы тематического моделирования текстов, а также подход отображения пространства слов в векторное пространство, дают хорошие результаты для нахождения слов схожих по тематике использования, а также по контекстам. Однако автоматическое раскрытие сокращений с помощью таких подходов, при использовании их напрямую, не достаточно точно. Полученные результаты можно применить для нахождения области использования неизвестных сокращений. А наиболее приемлемым способом для раскрытия сокращений остается поиск его по словарю сокращений. В таком случае возможна комбинация применения полученной программы и словаря сокращений. Например, можно использовать словарь [12]. По запросу расшифровки сокращения ресурс предложит список возможных вариантов полных форм. А построенные модели будут являться опорой для выбора конкретной расшифровки на основе контекстов данного сокращения.

Заключение

В работе была рассмотрена задача нахождения и раскрытия сокращений в корпусе документов РОМИП. Был построен словарь уникальных слов в начальной форме с их частотами встречаемости в корпусе. Рассмотрен способ нахождения сокращений с помощью выделения орфографических и семантических особенностей. При этом орфографические особенности были получены с помощью регулярных выражений на основе классификации сокращений. Для нахождения семантических особенностей была выдвинута гипотеза об информации содержащейся в частотных векторах соседей. В качестве меры информации принята энтропия. Изучен подход word2vec для анализа текста. Также исследована гетерогенная модель тематического моделирования LDA. Рассмотрен гибридный алгоритм для совместного использования моделей word2vec и LDA. Разработана программа-прототип для нахождения и раскрытия сокращений, успешно решающий первую задачу и дающий контекстную информацию для решения второй задачи.

Список литературы

[1] Blei D. Latent Dirichlet Alloacation

[2] Blei D., Chang J., Gerrish S., Wang C., Boyd-Graber J. L., Reading Tea Leaves: How Humans Interpret Topic Models // Advances in Neural Information Processing Systems 22,2009.P.288-296

[3] Heinrich G. Parameter estimation for text analysis

[4] Jenkins K. Deciphering Journal Abbreviations with JAbbr

[5] Manning C. Introduction to information retrieval

[6] Mikolov, T. Efficient Estimation of Word Representations in Vector Space [Electronic resourse]/ T. Mikolov, K.Chen, G. Corrado, J. Dean // arXiv.org – 2013 – URL: http://arxiv.org/pdf/1301.3781v3.pdf (date of access: 09.05.2016)

[7] Mikolov, T. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I. Sutskever, K.Chen, G. S Corrado, J. Dean // Advances in Neural Information Processing Systems. – 2013 – P. 3111-3119

[8] Moody C. http://www.slideshare.net/ChristopherMoody3/word2vec-lda-and-introducing-a-new-hybrid-algorithm-lda2vec-57135994

[9] Shannon C. A Mathematical Theory of Communication//Bell System Technical Journal 27. - 1948. 379-423

[10] Воронцов К.В. Вероятностное тематическое моделирование. http://www.machinelearning.ru/wiki/images/f/fb/Voron-ML-TopicModels.pdf

[11] Мильчин Ф.Э. Справочник издателя и автора. // М.:ОЛМА-Пресс,2003 — 800с. ISBN 5-224-04565-7

[12] Словарь сокращений. www.sokr.ru/

Приложение

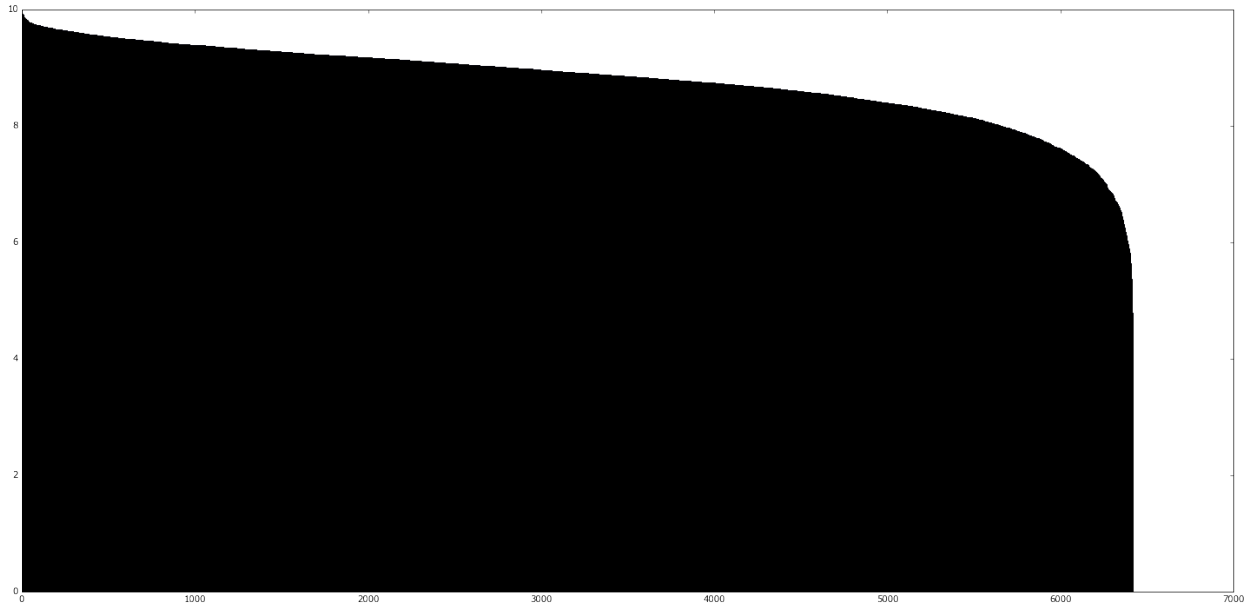

Рис. 8.3: Общее распределение энтропии.